文章编号：1003-0077（2017）00-0000-00

MonTTS：完全非自回归的实时、高保真蒙古语语音合成模型

刘瑞 [1] 康世胤 [2] 李劲东 [3] 飞龙 [1] 高光来 [1]

（1.内蒙古大学计算机学院 内蒙古自治区 呼和浩特 010021；2.虎牙科技有限公司，广州，511400；3.搜狗科技发展有限公司，北京，100000）

摘要：针对现有基于 Tacotron 模型的蒙古语语音合成系统存在的两个问题：1）难以实时合成；2）合成语音保真度较低，该文基于 FastSpeech2 模型提出了完全非自回归的实时、高保真蒙古语语音合成模型 MonTTS。为了提高 MonTTS 模型合成蒙古语语音的韵律自然度/保真度，根据蒙古语声学特点提出以下三点创新改进：1）使用蒙古文音素序列来表征蒙古文发音信息；2）提出音素级的声学调节器以学习长时韵律变化；3）提出基于蒙古语语音识别和自回归语音合成两种时长对齐方法。同时，该文构建了一个当前最大规模的蒙古语语音合成数据库：MonSpeech。实验结果表明 MonTTS 在韵律自然度方面的主观平均意见分数（Mean Opinion Score,MOS）达到 4.53，显著优于当前最优的基于 Tacotron 的蒙古语语音合成基线系统和基线 FastSpeech2 模型；MonTTS 合成实时率达 3.63×10^{-3}，满足实时高保真合成要求。最后，文中涉及的训练脚本和预训练模型全部开源（https://github.com/ttslr/MonTTS）。

关键词：蒙古语语音合成；非自回归声学建模；非自回归神经声码器；实时；高保真

中图分类号：TP391 **文献标识码：**A

MonTTS: A Real-time and High-fidelity Mongolian TTS Model with Complete Non-autoregressive Mechanism

LIU Rui[1] , KANG Shiyin[2] , LI Jingdong[3], BAO Feilong[1], and GAO Guanglai[1]

(1. Department of computer science，Inner Mongolia University， Hohhot 010021，China ; 2. Huya Inc, Guangzhou 511400, China；3. Sogou Inc，Beijing 100000，China)

Abstract ：Aiming at achieving real-time and high-fidelity speech generation for Mongolian Text-to-Speech (TTS), a FastSpeech2 based non-autoregressive Mongolian TTS system, termed MonTTS, is proposed. To improve the overall performance in terms of prosody naturalness/fidelity, MonTTS adopted three novel mechanisms: 1) Mongolian phoneme sequence was used to represent the Mongolian pronunciation; 2) phoneme-level variance adaptor was employed to learn the long-term prosody information; 3) two duration aligners, that are Mongolian speech recognition and Mongolian autoregressive TTS based models, were used to provide the duration supervise signal. Besides, we build a large-scale Mongolian TTS corpus, named MonSpeech. The experimental results show that our MonTTS outperforms the state-of-the-art Tacotron-based Mongolian TTS and standard FastSpeech2 baseline systems significantly, with real-time rate (RTF) of 3.63×10^{-3} and Mean Opinion Score (MOS) of 4.53, meeting the real-time and high-fidelity inference requirements. The training recipe and pretrained TTS models are freely available at https://github.com/ttslr/MonTTS.

Key words: Mongolian Text-to-Speech (TTS)，Non-autoregressive acoustic model，Non-autoregressive Neural vocoder，

收稿收稿日期：2021-10-22；**定稿日期**：20xx-xx-xx

基金项目：国家重点研发计划项目（2018YFE0122900），国家自然基金项目（61773224，62066033），内蒙古自然科学基金项目（2018MS06006），内蒙古自治区成果转化项目（CGZH2018125），内蒙古自治区应用技术研究与开发资金项目（2019GG372，2020GG0046）。

Real-time，High-fidelity

0 引言

语音合成主要将任意给定的文本转换为语音波形[1-2]。作为人工智能领域中的关键技术之一，它广泛应用于人机交互、泛娱乐、在线教育等领域[3]。

传统的语音合成方法主要包括基于波形拼接[3]和统计参数声学建模（如：隐马尔可夫模型[4-5]）的语音合成技术。随着深度学习技术的发展，基于深度神经网络结构的语音合成模型被广泛研究[6-8]。最终利用声码器将声学模型输出的语音参数转换为语音波形[9-10]，如 STRAIGHT[11] 和 WORLD[12]等。近年来，许多复杂的机器学习任务受益于强有力的深度神经网络模型，在性能上得到突破性的提升，也催生了端到端语音合成技术的研究[13]。端到端语音合成技术有效避免了传统多阶段建模导致的误差积累，同时简化了过多的人为假设，实现了媲美真实语音的合成效果。具体来说，主要包括两方面的研究突破：1）端到端声学建模；2）神经网络声码器。

对于声学建模研究，端到端声学建模主要采用“编码器-解码器”结构直接学习<文本，语音参数>对的对齐关系[14]，其中比较有代表性的是 Tacotron 模型[15]、Transformer 模型[16]及它们的多种变体[17-20]。以上模型在进行解码时，都是以上一时刻的输出作为下一时刻的输入进行声学参数的预测。这样的自回归解码结构极大限制了语音合成的实时性[21]，并不能充分利用目前高度发展的（如 GPU 等）并行计算硬件的计算资源。为了提高解码速度，研究人员进一步提出基于非自回归声学建模的语音合成模型[22]，如 FastSpeech[23]、FastSpeech2（s）[24]等。非自回归声学模型可以以给定文本为输入，并行输出全部声学参数序列，而不依赖于历史时刻解码得到的声学参数。

对于声码器研究，研究人员提出了基于神经网络的声码器来直接对语音样本点建模，如 WaveNet[25]，WaveRNN[26]等。神经声码器直接学习语音参数和语音波形采样点之间的映射关系，显著提高了合成语音的保真度[25]。但是基于 WaveNet 的声码器同样遵循自回归结构进行语音波形采样点的预测，这样的自回归生成过程耗时严重[27]。而语音重构的时间效率同样影响整个语音合成的实时性能。因此，为了加快神经网络声码器的语音生成速度，非自回归神经网络声码器逐渐受到广泛关注。如 Parallel WaveNet[28],WaveGlow[29],MelGAN[30]，HiFi-GAN[31]等。在合成语音高保真的同时，极大的提升了语音生成速度，能够达到实时语音生成。

当前，汉语和英语等主流语种的语音合成技术已发展较为成熟，低资源语言的语音合成逐渐受到越来越多研究人员的关注[32]。蒙古语隶属于阿尔泰语系蒙古语族蒙语支，它是蒙古语族中最著名且使用最广泛的语言[33]。在全世界范围内，使用人数大约有 600 万人[34]。同时，蒙古语也是中国内蒙古自治区的主体民族语言。因此，研究面向蒙古语的语音合成技术对于少数民族地区的教育、交通、通讯等领域具有重要意义。

为了开发和研究蒙古语语音合成系统，前人已经开展了大量的工作。文献[35-38]等结合蒙古语语言特点对基于波形拼接的传统语音合成方法进行研究。文献[39]提出了基于 HMM 声学模型的蒙古语语音合成的方法。文献[40]首次将深度学习技术引入蒙古语语音合成，使用基于 DNN 的声学模型代替 HMM 声学模型，进一步提升了蒙古语语音合成的整体表现；文献[41]实现了基于 Tacotron 的蒙古语语音合成系统。上述工作为蒙古语语音合成技术的研究奠定了坚实的基础。其中，基于端到端模型的蒙古语语音合成系统的合成语音的整体表现相较传统方法也获得了显著提升[41]。但是，基于 Tacotron 的端到端蒙古语语音合成系统在实时性和自然度两方面还有很多问题需要解决：1）现有端到端蒙古语语音合成模型采用自回归声学建模，依赖解码历史进行参数预测；2）语音重构模块使用 Griffin-Lim 算法等传统信号处理技术。传统算法进行语音重构时会不可避免的引入特征伪影[8]，限制了合成语音的音频保真度，导致合成语音与真人发音还有很大差距。因此，如何提升现有蒙古语语音合成系统的实时性和合成语音音频保真度，将是本文关注的重点。

如前所述，非自回归声学建模可以并行生成语音参数序列，与自回归声学建模相比，可以大大提升合成语音的效率。同时，非自回归神经声码器以语音参数为条件输入，可以直接对语音采样点进行精确预测，从而保证合成语音具有很好的音频保真度。

根据以上研究，为了解决蒙古语语音合成系统目前面临的实时性和音频保真度两个问题，本文首次提出了包括非自回归声学模型和非自回归神经声码器的完全非自回归蒙古语语音合成模型 MonTTS，其中非自回归模型基于当前最先进的 FastSpeech2[24]模型。但 FastSpeech2 中以语音帧为单位学习韵律变化的方式难以学习到蒙古语丰富的韵律变化，为了提高合成蒙古语语音的韵律自然度/保真度，我们面向蒙古语提出了以下三点创新性的改进：1）针对蒙古语文本表示，拉丁字符表示不足以表征蒙古语的发音信息，本文使用音素序列作为输入表示；2)针对蒙古语韵律建模，我们提出音素级别的基频、能量预测器，以更好的学习长时韵律变化；3）针对蒙古语时长建模，我们提出基于蒙古语语音识别和蒙古语自回归语音合成模型对训练数据的音素时长信息进行提取，为非自回归蒙古语时长预测提供精确的时长监督信息。对于非自回归神经声码器，为了快速生成高保真合成语音，我们选择当前最先进的基于生成对抗网络（GAN）的声码器：HiFi-GAN[31]，进行语音波形的重建。

为了确保基于数据驱动的端到端声学建模技术在蒙古语中得到充分训练，我们构建了当前最大规模（约 40 小时）的蒙古语语音合成语料库：MonSpeech。基于 MonSpeech 数据的一系列实验结果证明，本文提出的 MonTTS 模型在实时性和音频保真度两方面显著优于所有基线系统。

综上所述，本文主要贡献总结为如下几点：

- 本文提出了完全非自回归蒙古语语音合成模型 MonTTS，包括改进的非自回归声学建模和非自回归神经声码器。
- 本文针对蒙古语提出了三点创新的改进，包括音素序列的文本发音表示、音素级别的长时韵律建模、蒙古语音素时长监督提取等，在高效合成语音的同时有效保证了合成蒙古语语音的韵律自然度。
- 本文构建了目前最大规模（约 40 小时）的蒙古语语音合成语料库 MonSpeech，以尽可能满足基于数据驱动的端到端语音合成模型的训练数据需求。
- 本文首次针对非自回归蒙古语语音合成开展研究，填补了国内蒙古语语音合成研究的空白，本文工作也将对促进蒙古文智能信息处理和少数民族地区的人工智能技术发展贡献力量。

一系列主观和客观实验证明，本文的蒙古语语音合成模型 MonTTS 在音频保真度和实时性两方面均优于现有的蒙古语语音合成基线系统，并且可以为蒙古语上游语音交互系统提供基础服务。

论文结构安排如下：第二章介绍蒙古语文字及音系特点；第三章介绍蒙古语语音合成语料库 MonSpeech；第四章对本文提出的 MonTTS 系统的模型框架进行详细介绍；第五章展示详细的实验结果；最后对全文进行总结。

1 蒙古语语言特点

现行蒙古文拥有两种不同的书写系统[34]：西里尔蒙古文和传统蒙古文，传统蒙古文是一种拼音文字，本文的研究对象是传统蒙古文。

在文字表示方面，传统蒙古文形态丰富，其构词方式独特且复杂。汉语言文字在形态方面几乎不存在任何变化，其单词表示是由独立的字组成的，单词又进一步组成短语。蒙古文单词虽然也是由蒙古文字符直接拼接而成，但是与汉语相比其构词特点更加复杂，蒙古文单词是通过在词根或者词干后连接后缀构造而成。蒙古文单词可以拆分解构为多个组成部分：包括词根、构词后缀、构形后缀和结尾后缀等。

音系表示方面，音素是蒙古语发音的基本单元，蒙古语发音是由音素决定的，音素序列相比于字符序列能够更准确地表征发音信息。音节是由一个或几个音素组成的最小的语音片段，语音的节奏一般指语句中各音节的长短快慢。另外，音节单元和词干后缀一样，同样具有区别词义的功能。

本文将蒙古文拉丁序列表示中的每个拉丁单词称为单词（Word），将拉丁单词中的每个字母都称为字符（Character），音素序列中的每个音素单

元称为音素（Phoneme）。

2 蒙古语语音合成语料库 MonSpeech

MonSpeech 由内蒙古大学计算机学院授权，在内蒙古大学计算机学院标准录音室录制完成。文字抄本包含约 4 万条蒙古文语句，其中包含政治、商业、运动、娱乐等领域。该抄本覆盖了全部的蒙古文字母及丰富的单词组合情况。发音人为一名蒙古族女性专业蒙古语播音员，年龄 22 岁。最终录制数据总时长约 40 小时（其中平均每句话包含首尾静音段 0.3 秒），数据存储格式为：采样率 44.1 kHz,采样精度 16 bit。

表 1 MonSpeech 数据统计详情表

Category	Statistics	
Character	Total	2145828
	Mean	65
	Min	5
	Max	210
Phoneme	Total	2259159
	Mean	72
	Min	8
	Max	432
Word	Total	332500
	Mean	10
	Min	2
	Max	34
# Unique	59	
# Unique	38744	

MonSpeech 数据统计情况如表 1 所示。整个数据的蒙古文字符总数（Total）为 2145828 个，平均（Mean）每一句话包含 65 个字符，最短（Min）句子的字符个数是 5，最长（Max）句子的字符个数是 210。对于音素单元，MonSpeech 一共有 2259159 个音素，平均每句话包含 72 个音素，最短句子的音素个数是 8，最长句子的音素个数是 432。单词的总数、平均数、最大数量和最小数量分别为 332500、10、2 和 34。最终统计得到音素集合 59 个，词汇量 38744 个。另外，我们对数据中的句子时长进行统计，统计结果如图 1 所示。图中可以看到，大多数句子集中在 4 秒到 6 秒之间。由于 MonSpeech 中包含了大量蒙古文人名，因此 1 秒左右的语音比例达到了 1%。总体来说，句子时长服从正态分布。

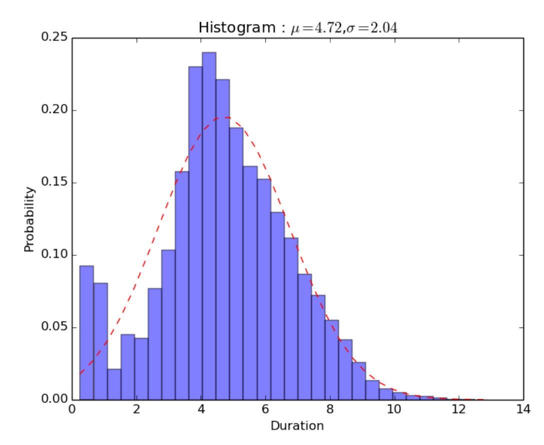


图 1 MonSpeech 句子时长统计直方图

3 MonTTS：完全非自回归的实时、高保真蒙古语语音合成模型

MonTTS 语音合成模型完全采用非自回归机制。具体包括非自回归蒙古语声学模型和非自回归神经声码器。如图 2a 所示，给定蒙古文句子，非自回归蒙古语声学模型以蒙古文句子的音素序列为输入，并行输出对应的 Mel 谱序列。非自回归声学声码器以句子的 Mel 谱序列为输入，并行输出语音采样点并最终输出合成蒙古语语音。下面将对这两部分内容做详细介绍。

3.1 非自回归声学模型

非自回归蒙古语声学模型以 FastSpeech2 为基础，具体结构包括蒙古文文本预处理（Mongolian Text Preprocessing）、蒙古文文本编码器（Mongolian Text Encoder）、蒙古语声学调节器（Mongolian Variance Adaptor）和蒙古语声学解码器（Mel Decoder）。其中，蒙古文文本预处理将输入的蒙古文句子转换为其音素表示，得到蒙古文音素序列（Mongolian Phoneme Sequence）。蒙古文文本编码器以蒙古文音素序列为输入，将其编码为高层的音素特征表示；蒙古语声学调节器内部的时长（duration）、基频（Pitch）、能量（Energy）预测器以音素向量为输入，分别预测出时长基频能量等声学信息并将其规整并附加到音素向量，得到调节后的隐含向量表示；最后声学解码器以隐含向量表示为输入对 Mel 谱进行并行预测。

需要注意的是，蒙古语文本编码器和声学解码器采用类似于 FastSpeech2[24] 中的结构。与 FastSpeech2 不同的是，我们的 MonTTS 针对蒙古

语的语言特性做出三点必要的创新改进：1）我们使用蒙古文预处理器将蒙古文文本转换为其音素序列表示。与拉丁字符序列相比，音素序列可以更好的表征蒙古文的发音信息；2）FastSpeech2 中的声学调节器只对帧级别的基频、能量信息进行预测。帧级别的声学信息不足以学习到音素级别的超音段韵律信息，从而不能很好的刻画长时变化的韵律结构。蒙古语属于黏着语，与汉语或英语相比，其发音具有很复杂的韵律变化[42]。为了更好的刻画蒙古语的长时韵律变化，在蒙古语声学调节器中，我们提出音素级别的基频、能量预测器，以学习蒙古文丰富的长时韵律变化；3）FastSpeech2 中的声学调节器在对英语句子的时长信息进行预测时，需要使用预提取的字符时间（语音帧的个数）信息提供精确的监督信号，而字符持续时间是一种与语种高度相关的信息。英语预提取时长信息在蒙古语场景下并不可用。为了对蒙古语时长预测器提供精确的监督信号，我们分别提出基于预训练蒙古语语音识别模型和蒙古语自回归语音合成模型两种方法来完成时长预提取，并将在实验部分对两者的效果差异进行比较。下面将对蒙古文文本预处理、蒙古语声学调节器和蒙古语时长预测器及相关损失函数进行详细介绍。

3.1.1 蒙古文文本预处理

传统蒙古文具有独特的黏着语特性，这为蒙古文文本处理带来很大挑战。具体来说，蒙古文字母在词中的表现形式变化不定，其显现形式在不同的上下文语境中会各不相同，因此导致蒙古文字母存在严重的形同音异现象。这种现象导致蒙古文文本数据中存在很多编码错误的字母。如前所述，本文的蒙古文文本预处理主要将蒙古文文本转换为其规范的音素序列表示。因此，蒙古文文本预处理包括编码校正，拉丁转换、文本正则化和字母转音素四个模块。首先，编码校正模块对输入蒙古文中的字符编码进行校正[43]，将编码错误的蒙古文字符转换为其正确的显现形式；之后，根据蒙古文拉丁字母对照表[44]，将校正后的蒙古文字符统一转换为拉丁表示形式。其次，针对蒙古文文本中经常出现的高频特殊字符（如：时间、日期、英文单词和阿拉伯数字等），设计了多达约 140 种[44]相应的正则表达式进行过滤。这 140 多种正则表达式基本覆盖了 95%以上的非蒙古文特殊字符情况[44]，可以准确的将不规范蒙古文文本转换为规范的蒙古文拉丁字符表示序列。之后我们使用基于联合对齐注意力机制的蒙古文字母转音素模型的蒙古文字母转音素（Grapheme to Phoneme，G2P）模型[45]将拉丁字符序列转换为其音素序列表示。该蒙古文字母转音素模型词错误率低至 6.2%，与汉语英语的前端文本处理模块[2,13]相比，已达到可用水平。因此，针对我们蒙古文文本正则化和蒙古文字母转音素模型的优秀表现。与汉语/英语等语音合成模型[2,13]类似，本文对前端文本处理模块中引入的不可避免的少量错误忽略不计，将得到的音素序列作为蒙古文文本

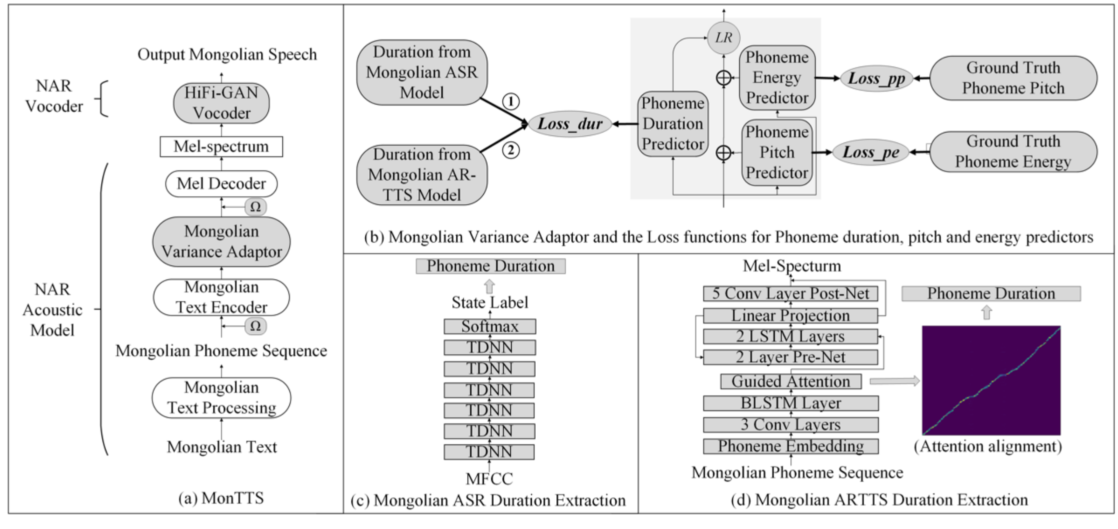


图 2　MonTTS 整体框架图，包括（a）模型结构；（b）音素级声学调节器内部结构及相应的损失函数；（c）基于蒙古语语音识别的对齐方法；（d）基于蒙古语自回归语音合成的对齐方法。

的表示序列。最后，文本编码器用来将音素序列进行语义信息编码，输出音素向量。

假设给定蒙古文文本X，蒙古文文本预处理将其转换为音素序列$W=\{W_1,W_2,\ldots,W_T\}$（T表示文本中音素的个数）。之后蒙古文文本编码器读取转换后的蒙古文音素序列W将其转换为深层次的音素向量$E'=\{E'_1,E'_2,\ldots,E'_T\}$。最后音素向量E'与位置向量Ω相加后得到新的音素向量$E=\{E_1,E_2,\ldots,E_T\}$被送入到蒙古语声学调节器对时长、基频、能量等信息进行预测。

3.1.2 蒙古语声学调节器

1）蒙古语时长预测器

蒙古语时长预测器以文本编码器输出的音素向量E为输入，对音素的持续时间信息$D=\{D_1,D_2,\ldots,D_T\}$（每个音素持续语音帧的数量）进行预测。其内部结构与 FastSpeech2 类似，包括 2 层搭配 ReLU 激活函数的 CNN 网络和 1 层全连接层。每一层 CNN 后都使用了正则化层（Layer Normalization，LN）和暂退层（Dropout Layer）来增加模型泛化性。

如图 2b 所示，在模型训练阶段，我们需要为蒙古语声学调节器中的时长预测器准备蒙古文音素的时长信息作为训练目标来计算时长损失函数:

$$Loss_{dur}=MSE(D,\widehat{D}) \quad (1)$$

其中$\widehat{D}=\{\widehat{D}_1,\widehat{D}_2,\ldots,\widehat{D}_T\}$表示预提取的音素时长信息。

参照汉语和英语等主流语言的最新进展，目前面向非自回归语音合成的时长预提取一般采用两种方案：1）使用预训练的语音识别模型作为对齐工具，对语音解码得到语言单位（如：字符、音素等）的时长信息；2）使用预训练的自回归语音合成模型，对语料库的文本进行前向计算，将得到的注意力对齐信息转换为时长信息。这两种方法在英语、汉语等语言表现出不错的效果，但是由于字符时长信息表现出高度的语言相关性，因此，英语或汉语的时长提取模型无法直接使用，而面向非自回归蒙古语语音合成的时长预提取也没有可用模型可以直接使用。

因此，如图 2b 所示，本文分别采用 ①大规模蒙古语语音识别数据下预训练的蒙古语语音识别模型以及 ②MonSpeech 下预训练的自回归蒙古语语音合成模型进行蒙古语音素时长信息的提取，作为时长预测器的训练目标来计算$Loss_{dur}$。下面将对这两种方法进行详细介绍。

① 基于蒙古语语音识别的对齐方法（ASRDur）：

如图 2c 所示，蒙古语语音识别模型以语音的梅尔倒频谱系数（Mel Frequency Cepstral Coefficients，MFCC）为输入，通过 6 层 TDNN 网络和 1 层 Softmax 输出层输出每个音素的状态标签[46-47]。最后，所需要的音素时长可以根据"状态标签-语音帧-音素"三者之间的对应关系转换得到[47]。

② 基于自回归蒙古语语音合成的对齐方法（ARTTSDur）：

如图 2d 所示，基于自回归声学建模的蒙古语语音合成模型以蒙古文音素表示为输入，通过"编码器-注意力-解码器"的模型结构对语音的 Mel 频谱参数进行预测。在训练阶段，编码器与解码器之间的注意力机制用来学习输入音素与输出语音帧之间的对齐关系。训练结束后，可以对任意输入蒙古文音素序列进行前向计算，得到该序列的注意力矩阵并从中解析出该输入序列中每个音素的持续时间。

基于自回归的蒙古语语音合成模型采用与 Tacotron2 类似的结构。编码器由 2 层 CNN 网络，1 层 BLSTM 网络组成。解码器由 2 层预处理 Pre-Net 网络，2 层 LSTM 网络，1 层线性层和 5 层基于 CNN 的后处理 Post-Net 网络组成。由于音素时长信息从注意力矩阵中解析得到，因此，注意力机制的选择对最终时长信息的精确性之间相关。为了更好地学习到呈对角线状态的注意力矩阵，与传统 Tacotron2 中的 location-aware attention[15]机制不同，本文采用 guided attention[48]机制对注意力矩阵进行对角线约束，从而可以实现更加精确的时长学习。

在实验部分，本文将对这两种方法提取的音素时长信息的准确性以及对非自回归声学建模的有效性进行详细验证和比较。

2）音素级基频和能量预测器

音素级基频和能量预测器以蒙古文文本编码器输出的音素向量E为输入，分别对音素级别的基频（Phoneme-level Pitch，PP）和能量（Phoneme-level Energy，PE）参数进行预测。

与 FastSpeech2 中基频和能量预测器对帧级别的基频、能量参数进行预测不同，本文的音素级基频、能量预测器对音素级别的基频、能量参

数进行预测。具体来说，我们先对语音求得每一帧的基频和能量参数，之后根据预提取的音素时长信息$\hat{D}=\{\hat{D}_1,\hat{D}_2,\dots,\hat{D}_T\}$对每一个音素的所有帧级别基频和能量参数求平均值，得到音素级别的基频和能量参数，分别记作$PP=\{PP_1,PP_2,\dots,PP_T\}$和$PE=\{PE_1,PE_2,\dots,PE_T\}$。

如图 2b 所示，在训练阶段，我们使用从训练数据中提取的真实的音素级基频和能量参数为目标来计算音素级的基频和能量损失函数，分别为$Loss_{pp}$和$Loss_{pe}$。

$$Loss_{pp}=MSE\left(PP,\widehat{PP}\right) \quad (2)$$

$$Loss_{pe}=MSE\left(PE,\widehat{PE}\right) \quad (3)$$

其中$\widehat{PP}$，$\widehat{PE}$表示音素级别的基频能量参数。

最后，时长规整器（Length Regulator，LR）根据时长预测器预测的字符时长$D=\{D_1,D_2,\dots,D_T\}$，将字符级别的文本特征向量$E=\{E_1,E_2,\dots,E_T\}$、基频向量$E_{PP}=\{E_{PP_1},E_{PP_2},\dots,E_{PP_T}\}$、能量向量$E_{PE}=\{E_{PE_1},E_{PE_2},\dots,E_{PE_T}\}$等信息，下采样为帧级别的联合特征向量$FE=\{Y_1,Y_2,\dots,Y_L\}$（$L$表示目标 Mel 谱的时长，即语音帧的数量），以与 Mel 谱进行长度匹配来并行预测梅尔频谱参数。Mel 谱解码器（Mel Decoder）读取联合特征向量FE来并行预测 Mel 谱$Y=\{Y_1,Y_2,\dots,Y_L\}$:

$$Y=MelDecoder(FE+\Omega) \quad (4)$$

其中，Ω与 3.1.1 节中的Ω相同，均表示位置编码。

综上所述，MonTTS 的非自回归蒙古语声学模型部分可以对蒙古文文本进行处理，将其实时转换为语音的 Mel 频谱特征表示。训练阶段的总损失函数$Loss$为$Loss_{mel}$，$Loss_{dur}$，$Loss_{cp}$，$Loss_{ce}$四个损失函数的总和。之后非自回归神经声码器将 Mel 谱特征实时生成语音波形。

3.1.2 非自回归神经声码器

非自回归神经声码器以语音的梅尔频谱$Y=\{Y_1,Y_2,\dots,Y_L\}$为输入，并行预测输出全部语音采样点$Z=\{Z_1,Z_2,\dots,Z_K\}$（K表示语音采样点的个数），最终输出语音波形。非自回归神经声码器可以对语音采样点进行并行生成，保证语音波形的实时生成。我们注意到非自回归与自回归神经声码器已经在汉语英语等语种中表现出优异的性能，但是当前蒙古语语音合成领域还只是停留在基于信号处理的声码器阶段，关于实时高保真的神经网络声码器的研究实现还处于空白阶段。

因此，为了填补这一空白，确保蒙古语语音合成又快又好，本文选择当前最优的基于 GAN 的声码器：HiFi-GAN[31]进行蒙古语语音波形的生成。

本文使用的 HiFi-GAN 与文献[31]具有相似的结构，包括一个生成器和两个判别器，两个判别器分别为多周期判别器和多尺度判别器。生成器是一个 CNN 网络，用来对 Mel 谱进行升采样，将长度L的 Mel 谱序列Y扩展到语音采样点长度K。多周期判别器通过观察输出音频菜不同周期的不同部分来捕获不同的隐式结构，多尺度判别器聚焦不同频率范围内的音频特征，从而保证语音信号的高保真生成。模型细节可见文献[31]。MonTTS 系统摒弃之前使用的 Griffin-Lim 语音重构算法，首次使用蒙古语语音合成数据成功训练得到高质量的蒙古语 HiFi-GAN 声码器，可以在实时 Mel 频谱参数预测的基础上，实时合成高保真的蒙古语语音。我们将在下一章的实验部分对 MonTTS 系统的性能进行验证。

4 实验

4.1 实验数据

蒙古语语音合成模型训练数据：我们基于本文构建的 MonSpeech 数据集进行语音合成的训练。如第二章中介绍，MonSpeech 包含约 40 小时的单说话人蒙古语语音及其对应约 4 万句文字抄本。我们将数据按照 8：1：1 的比例划分为训练集、验证集和测试集。

蒙古语语音识别模型预训练数据：针对基于蒙古语语音识别的时长预提取方法，我们使用内蒙古大学计算机学院所有的约 1500 小时的多说话人蒙古语语音识别标准数据[46]进行语音识别模型的训练。文献[46]首次使用该数据进行蒙古语语音识别实验，请阅读文献[46] 了解该数据更多信息。

4.2 对比实验设计

为了验证本文提出的 MonTTS 在解码效率和语音音质两方面的表现，本文一共构建了 6 个系统：

（1）Tacotron2（GL）：该系统使用基于自回

归机制的 Tacotron2 语音合成模型进行 Mel 谱参数预测，之后使用 Griffin-Lim 算法进行语音重构；

（2）Tacotron2（HiFiGAN）：该系统同样使用 Tacotron2 模型预测 Mel 谱，与第一个系统不同的是，使用 HiFiGAN 声码器进行语音的生成。

（3）FastSpeech2+ASRDur（HiFiGAN）：该系统采用 FastSpeech2 模型进行 Mel 预测，使用 HiFiGAN 声码器进行语音生成。其中，时长预测器的训练目标由蒙古语语音识别模型提供。ASRDur 表示基于蒙古语语音识别模型的时长预提取方法。

（4）FastSpeech2+ARTTSDur(HiFiGAN)：该系统采用 FastSpeech2 模型进行 Mel 预测，使用 HiFiGAN 声码器进行语音生成。其中，时长预测器的训练目标由自回归蒙古语语音合成模型提供。ARTTSDur 表示基于自回归蒙古语语音合成模型的时长预提取方法。

（5）MonTTS+ASRDur(HiFiGAN)：该系统采用本文提出的 MonTTS 模型进行 Mel 预测，使用 HiFiGAN 声码器进行语音生成。与 FastSpeech2+ASRDur 类似，时长预测器的训练目标由蒙古语语音识别模型提供。与（3）和（4）相比，MonTTS 在蒙古语声学调节器中使用字符级的基频和能量预测器。

（6）MonTTS+ARTTSDur(HiFiGAN)：该系统同样采用 MonTTS 模型和 HiFiGAN 声码器。其中，时长预测器的训练目标由自回归蒙古语语音合成模型提供。

4. 4 实验设置

MonTTS 的模型参数与 FastSpeech2[24]相似。文本编码器和声学解码器均包含 4 层 FFT 模块，音素向量和内部的隐层向量都是 256 维。我们将语音数据重采样到 22.05kHz 并采用帧长 50ms，帧移 12.5ms 提取 80 维的 Mel 谱参数。Pitch 和 Energy 也使用与 FastSpeech2[24]相同的参数配置计算。Dropout 比率设置为 0.5。 batch size 大小设置为 32。我们使用与 FastSpeech2 相似的学习率动态调整方法训练模型 200k 步。其余 Tacotron2 和 FastSpeech2 模型同样训练 200k。对于 HiFi-GAN 声码器，我们先训练生成器 100k 步，之后联合训练生成器和判别器 300k。ASRDur 方法中 6 层 TDNN 的上下文语音帧扩展配置为[-1,0,1], [-1,0,1,2]， [-3,0,3]， [-3,0,3]， [-3,0,3]， [-6,-3,0] 。ARTTSDur 方法中的编码器解码器参数配置与 Tacotron2 相同。

4. 4 实验结果

4.4.1 蒙古文文本表示比较

我们首先基于 Tacotron2（GL）模型对蒙古文的文本表示方法进行比较。我们分别使用拉丁字符表示和音素表示进行模型的训练，并进行主观听力测试比较二者合成语音的质量。我们从测试集中随机选取 50 句蒙古文文本并分别使用字符和音素序列表示进行语音合成，之后邀请 10 位蒙古族青年学生对 100 句合成语音进行 MOS[44]打分（5-优秀，4-良好，3-可接受，2-一般，1-很差）。实验结果如图 3 所示，音素序列表示的 MOS 分数为 3.98，显著优于字符序列的分值 3.82。表明音素序列与蒙古文的发音信息直接相关，可以合成自然度更高的语音。之后的实验中，所有的蒙古语语音合成系统均以音素序列作为输入。

4.4.2 蒙古文时长对齐方法比较

我们首先对基于蒙古语语音识别和自回归蒙古语语音合成两种音素时长对齐方法的准确度进行比较。我们从测试集中随机选取 50 句蒙古语语音及其对应音素序列，使用 Praat 软件[1]进行音素时长标注。之后，我们分别使用两种对齐方法得到的音素时长和标注的真实音素时长计算时长准确率（phoneme duration accuracy）[24]。实验结果如表 2 所示，从表中可以看到，语音识别对齐方法相比自回归语音合成对齐方法可以得到更精确的时长信息。分析原因可能有以下两点：1）蒙古语语音识别模型基于 1500 小时的大规模多说话人训练数据训练得到，模型具有很好的泛化性，可以得到精确的“状态标签-语音帧-音素”对应关系；2）基于自回归语音合成对齐方法中，注意力机制的选择是能否得到精确对齐关系的关键之一，对角线指导的 guided attention 还没有体现出注意力对齐的单调特性，可能导致对齐信息不够精确。

[1] https://www.fon.hum.uva.nl/praat/

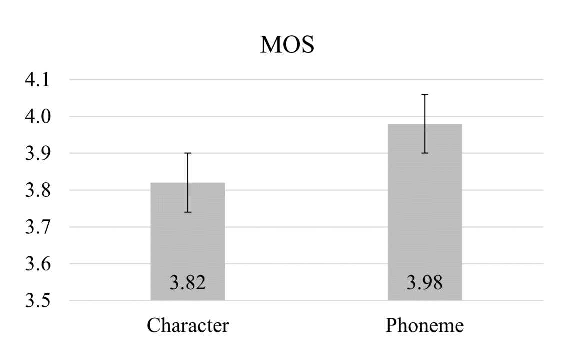


图 3　不同蒙古文文本表示的 MOS 分数（95%置信度）

表 2　ASRDur 和 ARTTSDur 两种方法的对齐精度比较

Method	Δ (ms)
ASRDur	**14.02**
ARTTSDur	18.13

4.4.3 蒙古语语音合成韵律自然度比较

本节我们将从客观实验和主观实验两个方面对不同系统在合成语音音质方面的性能进行详细比较。

对于客观实验，我们选择 Mel Cepstral Distortion(MCD)[49]以及 pitch 和 energy 的均方误差（Mean Absolute Error，MAE）三个指标进行语音质量的测试。我们从测试集随机选取 50 句文本并使用不同系统进行语音合成，之后分别从合成语音和真实语音中提取出 mel、pitch 和 energy 三种声学特征并且计算上述三个指标。在指标计算时我们使用 Dynamic time warping （DTW）算法[50]将合成语音和生成语音进行对齐。实验结果如表 3 所示，可以发现：1）所有的非自回归语音合成模型 FastSpeech2 和 MonTTS 都明显优于 Tacotron2 模型；2）本文提出的 MonTTS 模型与 FastSpeech2 相比可以合成更接近真实语音的语音参数，从而合成更加自然的语音，也证明了我们的音素级别声学调节器可以更好的学习蒙古语的长时韵律特征从而生成韵律更自然的语音；3）与 MonTTS+ARTTSDur 相比，MonTTS+ASRDur 输出的语音参数明显较优。从另一角度证明语音识别对齐方法可以提供更加准确的时长监督，有利于更加自然的语音生成；4）本文基于 MonSpeech 训练的 HiFi-GAN 声码器显著优于传统的 Griffin-Lim 算法，可以得到高质量的合成语音。

对于主观实验，我们进行主观 MOS 听力测试。测试样本选择与客观实验相同，我们邀请 15 位蒙古族青年学生对所有合成语音和对应的真实语音进行 MOS 打分。实验结果如图 4 所示，从图中可以看出本文提出的 MonTTS 系统搭配我们首次基于 MonSpeech 数据训练得到的 HiFi-GAN 声码器，可以输出高保真的合成语音，合成语音获得了接近 4.53 的 MOS 分数，显著优于所有基线系统，并且与真实语音的 MOS 分数基本相当。

上述客观实验和主观实验的实验结果充分证明本文提出的 MonTTS 系统在合成语音音质方面的强大性能。下一节我们将比较不同系统在合成效率方面的表现。

表 3　针对语音自然度的客观实验结果

Method	MCD	MAE	
		Pitch	Energy
Tacotron2 (GL)	8.53	20.73	0.552
Tacotron2 (HiFi-GAN)	8.14	18.90	0.502
FastSpeech2+ARTTSDur (HiFi-GAN)	7.74	18.89	0.483
FastSpeech2+ASRDur (HiFi-GAN)	7.68	18.86	0.479
MonTTS+ARTTSDur (HiFi-GAN)	7.53	18.41	0.462
MonTTS+ASRDur (HiFi-GAN)	7.38	18.32	0.438

4.4.4 蒙古语语音合成效率比较

我们同样采用上一节的 50 句测试集进行语音合成速度测试。我们使用不同系统对 50 句测试集进行 10 次语音合成，统计每次合成所需要的时间。之后以 50 测试集对应真实语音的时间为参照，计算语音合成实时率（Real-time factor，RTF）[24]。实验结果如表 4 所示，从表中我们发现 1)MonTTS（HiFi-GAN）与 Tacotron2（HiFi-GAN）相比，实时率显著提升，说明非自回归声学模型在合成效率上显著优于自回归 Tacotron2 结构；2）Tacotron2（HiFi-GAN）与 Tacotron2（GL）相比，说明本文训练得到的 HiFi-GAN 声码器同样凭借其非自回归的快速波形生成能力在合成效率上表现出优秀的性能。本文提出的 MonTTS（HiFi-GAN）的合成实时率达到了 3.63×10^{-3}，已经达到实时合成，可以很好的满足实际应用需求。

综上所述，本文提出的 MonTTS 模型在合成语音音质和合成效率两方面均表现出优异的性能，显著优于所有基线系统。MonTTS 实现了第一个全非自回归的实时、高保真蒙古语语音合成系统。

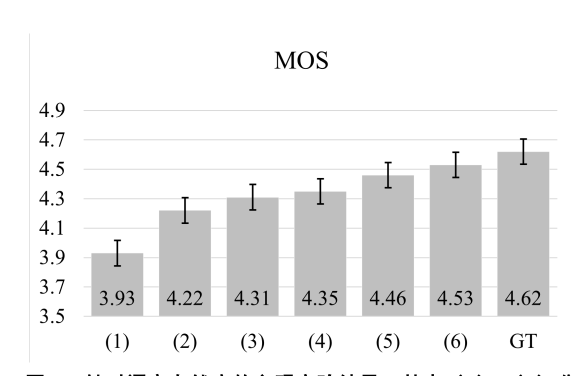


图 4 针对语音自然度的主观实验结果，其中（1）-（6）分别表示 4.2 节中构建的 6 个对比系统，由于空间限制仅使用其数字编号代表系统名称

表 4 不同系统的合成实时率比较

Method	RTF
Tacotron2 (GL)	9.13×10^{-1}
Tacotron2(HiFi-GAN)	9.01×10^{-1}
MonTTS(HiFi-GAN)	$\mathbf{3.63 \times 10^{-3}}$

5 总结

本文提出首个完全非自回归机制的实时、高保真蒙古语语音合成系统 MonTTS。基于当前先进的 FastSpeech2 并针对蒙古文文本表示、蒙古文韵律建模和蒙古文时长建模提出音素表示、音素级基频和能量预测器以及基于蒙古语语音识别和自回归蒙古语语音合成的时长对齐方法。实验结果表明，本文提出的 MonTTS 在语音质量和合成效率两方面优于所有基线系统，达到高保真语音的实时合成，可以为上游蒙古语语音交互系统提供全新的技术服务。本文实验仅使用单一女性说话人语料进行实验，为了更好的验证模型在不同说话人的效果，未来工作将收集整理更多的说话人数据（包括不同年龄段的男性和女性说话人等）对 MonTTS 模型的有效性进行验证。更进一步，未来研究将对该模型进行扩展，实现高质量的多说话人和多情感的蒙古语语音生成能力。

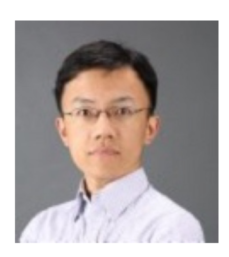
参考文献

[1] Taylor P. Text-to-speech synthesis[M]. Cambridge university press, 2009.
[2] Zen H, Tokuda K, Black A W. Statistical parametric speech synthesis[J]. speech communication, 2009, 51(11): 1039-1064.
[3] Hunt A J, Black A W. Unit selection in a concatenative speech synthesis system using a large speech database[C]//1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. IEEE, 1996, 1: 373-376.
[4] Tokuda K, Yoshimura T, Masuko T, et al. Speech parameter generation algorithms for HMM-based speech synthesis[C]//2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100). IEEE, 2000, 3: 1315-1318.
[5] Yamagishi J, Ling Z, King S. Robustness of HMM-based speech synthesis[J]. 2008.
[6] Ze H, Senior A, Schuster M. Statistical parametric speech synthesis using deep neural networks[C]//2013 ieee international conference on acoustics, speech and signal processing. IEEE, 2013: 7962-7966.
[7] Wu Z, Swietojanski P, Veaux C, et al. A study of speaker adaptation for DNN-based speech synthesis[C]//Sixteenth Annual Conference of the International Speech Communication Association. 2015.
[8] Zangar I, Mnasri Z, Colotte V, et al. Duration modeling using DNN for Arabic speech synthesis[C]//9th International Conference on Speech Prosody. 2018.
[9] Griffin D, Lim J. Signal estimation from modified short-time Fourier transform[J]. IEEE Transactions on acoustics, speech, and signal processing, 1984, 32(2): 236-243.
[10] Agiomyrgiannakis Y. Vocaine the vocoder and applications in speech synthesis[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 4230-4234.
[11] Kawahara H. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds[J]. Acoustical science and technology, 2006, 27(6): 349-353.
[12] Morise M, Yokomori F, Ozawa K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications[J]. IEICE TRANSACTIONS on Information and Systems, 2016, 99(7): 1877-1884.
[13] Wang W, Xu S, Xu B. First Step Towards End-to-End Parametric TTS Synthesis: Generating Spectral Parameters with Neural Attention[C]//Interspeech. 2016: 2243-2247.
[14] Wang Y, Skerry-Ryan R J, Stanton D, et al. Tacotron: Towards end-to-end speech synthesis[J]. arXiv preprint arXiv:1703.10135, 2017.
[15] Shen J, Pang R, Weiss R J, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 4779-4783.
[16] Li N, Liu S, Liu Y, et al. Neural speech synthesis with transformer network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 6706-6713.
[17] Liu R, Sisman B, lai Gao G, et al. Expressive tts training with frame and style reconstruction loss[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021.
[18] Liu R, Sisman B, Li J, et al. Teacher-student training for robust tacotron-based tts[C]//ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2020: 6274-6278.
[19] Liu R, Sisman B, Bao F, et al. Modeling prosodic phrasing with multi-task learning in tacotron-based TTS[J]. IEEE Signal Processing Letters, 2020, 27: 1470-1474.
[20] Liu R, Sisman B, Li H. Graphspeech: Syntax-aware graph attention network for neural speech synthesis[C]//ICASSP

2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6059-6063.
[21] Elias I, Zen H, Shen J, et al. Parallel Tacotron 2: A Non-Autoregressive Neural TTS Model with Differentiable Duration Modeling[J]. arXiv preprint arXiv:2103.14574, 2021.
[22] Liu R, Sisman B, Lin Y, et al. FastTalker: A neural text-to-speech architecture with shallow and group autoregression[J]. Neural Networks, 2021, 141: 306-314.
[23] Ren Y, Ruan Y, Tan X, et al. FastSpeech: fast, robust and controllable text to speech[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 3171-3180.
[24] Ren Y, Hu C, Tan X, et al. Fastspeech 2: Fast and high-quality end-to-end text to speech[J]. arXiv preprint arXiv:2006.04558, 2020.
[25] van den Oord A, Dieleman S, Zen H, et al. WaveNet: A Generative Model for Raw Audio[C]//9th ISCA Speech Synthesis Workshop. 125-125.
[26] Kalchbrenner N, Elsen E, Simonyan K, et al. Efficient neural audio synthesis[C]//International Conference on Machine Learning. PMLR, 2018: 2410-2419.
[27] Paine T L, Khorrami P, Chang S, et al. Fast wavenet generation algorithm[J]. arXiv preprint arXiv:1611.09482, 2016.
[28] Oord A, Li Y, Babuschkin I, et al. Parallel wavenet: Fast high-fidelity speech synthesis[C]//International conference on machine learning. PMLR, 2018: 3918-3926.
[29] Prenger R, Valle R, Catanzaro B. Waveglow: A flow-based generative network for speech synthesis[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 3617-3621.
[30] Kumar K, Kumar R, de Boissiere T, et al. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis[J]. Advances in Neural Information Processing Systems, 2019, 32.
[31] Kong J, Kim J, Bae J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis[J]. Advances in Neural Information Processing Systems, 2020, 33.
[32] Xu J, Tan X, Ren Y, et al. Lrspeech: Extremely low-resource speech synthesis and recognition[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 2802-2812.
[33] Liu R, Sisman B, Bao F, et al. Exploiting morphological and phonological features to improve prosodic phrasing for mongolian speech synthesis[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 29: 274-285.
[34] Bulag U E. Mongolian ethnicity and linguistic anxiety in China[J]. American Anthropologist, 2003, 105(4): 753-763.
[35] Davaatsagaan M, Paliwal K. Diphone-based concatenative speech synthesis system for mongolian[C]//Proceedings of the International MultiConference of Engineers and Computer Scientists. 2008, 1: 19-21.
[36] Lai G G, Min H, Qin Z S. The research and implementation of Mongolian text to speech system[C]//6th International Conference on Signal Processing, 2002. IEEE, 2002, 1: 472-475.
[37] Qi B. On some problems about the text in mongolian speech synthesis[C]//2017 International Conference on Asian Language Processing (IALP). IEEE, 2017: 48-51.
[38] HE M, GAO G, ZHAO S. The Research and Implementation of Mongolian Text to Speech System[J]. Acta Scientiarum Naturalium Universitatis Neimongol, 2004, 1.
[39] Zhao J, Gao G, Bao F, et al. Research on hmm-based mongolian speech synthesis[J]. Computer Science, 2014, 41(1): 80-104.
[40] Liu R, Bao F, Gao G, et al. Mongolian text-to-speech system based on deep neural network[C]//National conference on man-machine speech communication. Springer, Singapore, 2017: 99-108.
[41] Li J, Zhang H, Liu R, et al. End-to-end mongolian text-to-speech system[C]//2018 11th international symposium on chinese spoken language processing (ISCSLP). IEEE, 2018: 483-487.
[42] Liu R, Bao F, Gao G, et al. Improving Mongolian Phrase Break Prediction by Using Syllable and Morphological Embeddings with BiLSTM Model[C]//Interspeech. 2018: 57-61.
[43] Liu R, Bao F L, Gao G, et al. Phonologically aware BiLSTM model for mongolian phrase break prediction with attention mechanism[C]//Pacific Rim International Conference on Artificial Intelligence. Springer, Cham, 2018: 217-231.
[44] Liu R, Bao F, Gao G. Building mongolian tts front-end with encoder-decoder model by using bridge method and multi-view features[C]//International Conference on Neural Information Processing. Springer, Cham, 2019: 642-651.
[45] Wang Y, Bao F, Zhang H, et al. Joint Alignment Learning-Attention Based Model for Grapheme-to-Phoneme Conversion[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 7788-7792.
[46] Wang Y, Bao F, Zhang H, et al. Research on Mongolian speech recognition based on FSMN[C]//National CCF Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2017: 243-254.
[47] McAuliffe M, Socolof M, Mihuc S, et al. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi[C]//Interspeech. 2017, 2017: 498-502.
[48] Tachibana H, Uenoyama K, Aihara S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 4784-4788.
[49] Kominek J, Schultz T, Black A W. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion[C]//Spoken Languages Technologies for Under-Resourced Languages. 2008.
[50] Senin P. Dynamic time warping algorithm review[J]. Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, 2008, 855(1-23): 40.



刘瑞（1994–），博士，主要研究领域为机器学习、基于深度学习的语音合成和自然语言处理等。
E-mail：liurui_imu@163.com



康世胤（1984—），博士，主要研究领域为基于深度学习的语音合成、语音转换等。
E-mail：kangshiyin@huya.com



高光来（1964—），通信作者，硕士，教授，主要研究领域为模式识别、机器学习、深度学习、蒙古文信息处理等。
E-mail：csggl@imu.edu.cn